Machine Learning in Cybersecurity

Dominic Lanzante, Sai Santhoshi Arcot, Deekshith Rangineni, Aaliya Moore

University of Nebraska at Omaha

CYBR: 4360

Sayonnha Mandal

12/08/2024

Table of Contents

1. Introduction

When the reader/audience is asked to imagine what machine, learning is in cybersecurity there are a lot of blank stares and open-ended questions. In short, Machine Learning (ML) in cybersecurity (cyber) refers to the use of complex algorithms and statistical models further integration into our now existing computer systems. Think of it as an extra blanket for when you are tucking yourself in at night to make sure you don't get cold or overheated, supposedly this helps to better regulate other things in an organizational security posture. Machine learning is a new anonymous way of detecting, analyzing, and responding to cyber threats without any explicit application, program, or website. In cyber-ML can sift through massive amounts of data to identify patterns, predict attacks, predict potential vectors, as well as make decisions based on the observed network. This is also aiding in the assistance of safeguarding information but also keeping the infrastructure safe as well.

There are many key and primary roles of machine learning. The authors will be going over key applications of ML in cyber such as threat detection, malware detection, phishing and fraud, as well as anomaly detection.

2. The Role of Machine Learning in Cyber

ML in cybersecurity has three main objectives as stated above and they are threat detection , prevention, and response. ML addresses the ever-increasing challenges of sophisticated and volume of traditional cyber threats. With ML it is made seamless to detect upcoming cyber threats but to also create trends in your org that can show you when a cyber-attack might happen next and hope to best prevent it. One of the primary functions of ML in cyber is anomaly detection as mentioned above. Now taking this into account we as a reader must also come to realize that these MLs are taking in massive amounts of data, such as user behaviors, interactions, network traffic, or even system logs at that time to learn patterns and learn the behaviors of the system. When any unusual activity occurs outside the normal operating procedures set by the ML then it gets flagged and put into a certain category for viewing or will generate alerts. ML can also play a key role in identifying vulnerabilities that we thought once to be unrecognizable or to be something that could never be uncovered. These are called zero-day attacks. The final aspect and key role or responsibility that we might uncover with ML is it's scalability, meaning orgs can throw massive amounts of data at it and get generated threats that once also thought to be overlooked by the human eye.

3. Annotated Bibliography

This is going to be an annotation of each source finding that the authors found. This is going to include a summary, the formal citation, evaluation, and reflections of what the authors thought of the Scholar Articles.

3.1 Digital Threats: Research and Practice

Machine Learning is a crucial technology being developed for both current and future information systems. Countless other fields are benefiting from ML but why isn't Cyber? This is a key takeaway from this article that distinguishes it from the rest because it highlights a significant gap that there is between where the current technology is and where it needs to be to oversee such capabilities. The authors also believe that if the strengths and weaknesses of machine learning are not clearly defined by a mass audience or by the common corporations then it will be just like the wild west. This article aims to offer a wider more comprehensive view of ML and cyber as a whole and how each individual aspect has been operating for some time, but they now have the chance to finally meet and see what the possibilities are. The author also aimed to mention that this article was for more practical real-world deployment of ML and aims to explore multiple scenarios for which ML can be applicable today. This is crucial for the advancement of cybersecurity and ML. This is because without knowing the threat landscape and knowing possible avenues for attacks then you will not know where to evenly deploy your ML to pick up on anomalies. The authors have evaluated this article and deemed it fit to be in our project review because of its relevance and how it can be related to others as well as other organizations. We also believe that with ML certain cyber processes can be streamlined. As the authors stated above it is a very professional article, and the best take away from this article is that we have much to learn and a lot of testing to be done before we can turn this loose into a more real-world setting.

3.2 Testing Cyber-Physical Systems via Evolutionary Algorithms and ML

The following article was evaluated by the authors and the key takeaway and what really stood out about this article was that it emphasizes the challenges of cyber-physical systems and tries to ensure reliability and safety of cyber-physical systems. This study aimed to explore different methodologies and how certain methodologies such as model-based design, model-based engineering, and segmentation. Each individual aspect has its own unique category and unique challenges that are faced when thrown into a ML language. Machine learning languages can now automatically generate test cases and adapt these complex systems into already streamlined pipelines. The authors believe that with the combination of evolutionary algorithms from the ML it aims to enhance the security posture of an organization to provide better testing, coverage, and efficiency. The authors found this article to be highly credible and believe that the article just did not go deep enough down the rabbit hole pre-se. The information is aligned with another research. The authors just believe that in other articles proceeding with this one it should have added in the potential for large language models. The author's final perspective on this research article is that the article is especially useful just because there are more useful articles provided within this research paper than this one. As the author states above, there is just too little practical exercise being done and we as security professionals and researchers need to be able to recreate what we are writing and talking about. The authors would love to have seen an integration into ChatGPT or even copilot and this would have been at the top of our list.

3.3 Detecting Insider Threats in Cybersecurity Using Machine Learning and Deep Learning Techniques

In the cyber landscape and ever-evolving threat model sometimes these threats are generated from their own employees. These insider risks usually are from contractors or internal employees that have been done wrong one too many times. In “Detecting Insider Threats in Cybersecurity Using Machine Learning and Deep Learning Techniques” the researchers present a comprehensive way to identify insider hazards by making use of both deep learning and ML. Lee and Thompson had done work in the later quarter of 2021 using support vector machines (SVM), Long Short-term Memory (LSTM), and they also played around with the fact that with SVM’s, ML, and deep learning the organization can achieve and boast a 96.9% insider threat detection before it happens. This means that you are more likely to get into a car accident than to get past these learning models that detect system logs, user patterns, behaviors, reCAPTCHA, and more. These deep learning and machine learning aspects are coming together to create uniquely amazing results. In addition to the 96% insider threat detection rate, it has a less than 2% false positives rate from the total. In conclusion the authors believe that not only for insider detection but for external detection such as phishing, vishing, tailgating, any other type of cyber-attack, deep learning and machine learning are the best way to go. With the ever-evolving space that we are going into that is cyber we can use these two great technologies to better predict and analyze patterns and trends better. Finally, for the best real-time detection and threat hunting model use both deep learning and machine learning methodologies.

3.4 Harnessing the Speed and Accuracy of Machine Learning to Advance Cybersecurity

Traditional security measures, which often rely on signature-based detection and predefined rule sets. To address difficult challenges such as the one mentioned above as well as increasingly sophisticated attacks, organizations are turning to ML and DL as powerful tools to enhance the speed of IDS and IPS detection and prevention. ML algorithms can learn from vast amounts of different data that they can pull in from these log aggregators. The increasing ability of ML models to analyze patterns in real-time and adapt to previously unseen attacks such as all the ones mentioned above. This article sets itself apart because it tries to explore automating these threats as well as making the most of these manual threat detection and prevention tasks. Through the integration of ML organizations, we can build the best security posture possible. This article tends to set itself apart from the other ones because it tries to pull in distinct aspects from unsupervised learning, deep learning, and another not explored topic of reinforcement learning. The authors also agree that reinforcement learning and unsupervised learning both excel uniquely at different things. Un supervised learning excels at discovering hidden structures and patterns that data without labels can get and alternatively reinforcement learning is centered around decision-making and learning in particularly dynamic environments. Together they both broaden the scope of machine learning by providing highly flexible methodologies for dealing with unlabeled data and dynamic decision-making environments.

3.5 Blockage of Phishing Attacks Through Machine Learning Classification Techniques and Fine Tuning its Accuracy

Hacking attempts against internet users most likely follow with more of a phishing assault. Those individuals that do use phishing techniques use a range of social engineering tactics, and they often tweak their plan based on who they are massaging and how they are messaging back. To fool individuals they craft emails, messages, social media accounts and DM's, and other secret communication avenues that seem genuine but are just a rouse. The best way that the authors suggest is mitigating the major risks caused by spam emails. Most of the major email companies that you have heard of most likely have already incorporated convolutional neural network (CNN) built into their filtering systems. The authors have been notified of a new methodology called the K-fold cross-validation approach. Now this might sound like anything right now, but the author will break that term down and how this is different from traditional ML. This article used the main new methodology of the k-fold cross validation approach, and this methodology came with improved generalization, more reliable metrics, efficient data manipulation, better hyperparameter tuning and finally reduction evaluation resulting in a more consistent and less volatile measure of mode performance. The authors believe that this article provided further research and insight into what really needs to be looked at and done with ML in regard to CNN.

3.6 Denial-of Service (DoS) Attack Detection Using Edge Machine Learning

We all know traditional Denial-of-service (DoS) attacks and how they are intended to control networks and how with things being even more connected than they ever have been with IoT devices it makes things such as DoS attacks very hard to mitigate or even see coming as things such as DoS attacks usually end up in zero-day attacks. DoS attacks mitigation traditionally have been implemented on cloud models but usually that is not always possible to keep your cyber defense and IPS/IDS on a cloud-based model. Usually, organizations keep the cloud model and eat the computational power and cost in other fees. This article takes all of the cloud costs and takes a new spin to see what the computational power, load, balance and how much it would cost to actually make the DoS mitigation system a physical part of the network making it an edge device that sat where most IPS/IDS would lie as well. With this it gives the ML a unique advantage as it sees all the traffic as it is first coming into the network so that even before the firewall and allow list or block-list can say yes or not the ML will decide based on other human and organizational behaviors and then make the best decision as to what category the incident should fall under. This article gave a unique perspective because it took what the authors knew about DoS and enhanced the knowledge by incorporating a new aspect to think about as well as new methodologies such as the development and deployment phases to make sure that the ML is picking up all the proper information. The authors have deemed this article to be a particularly credible source to make claims and highlights from. This article can be used in multiple different disciplines and settings and will be hoping to be able to use this in the future. In the end the article suggested that ML be trained in set data points and set analysis so that it can have better data to go off. It was reported that the models trained in real-time had higher accuracy but lacked features for analysis.

3.7 Cybersecurity data science: an overview from machine learning perspective

Due to the increasing digitization and ever-growing threat zone, it is getting increasingly difficult to know what are false positives? What are true data breaches? Where does the real start and the fake end? Take a stat from this article for instance, "in 2019, there are more than 900 million malicious executables known to the security community, and this number is likely to grow" (). ML is an ever growing and evolving item on the docket as well. In regard to data science we can tap into even more realms and places that before were thought to be unattainable. As mentioned above and in previous articles ML is making smarter and smarter decisions as the days go by. This means as we are also making mistakes as humans trying to get our incidents managed the

ML applied with data driven science behind it could have already alerted on six other incidents that previous the user/author would be still trying to close and finalize the last ticket. It gets even better when the organization has ML set up with their other apps to better respond and figure out better ways the ML can be deployed org wide. In the authors opinion cybersecurity and ML in data science have been thoroughly documented on the academia side but the real-world application of most of these methodologies and approaches have not been documented by the organizations or any governing body. We would love to see some collaboration going on there.

3.8 A Comprehensive Survey: Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions

Now we have looked at the effectiveness of machine learning, deep learning, and reinforcement learning. These play an absolute crucial role in mitigating everyday threats and anomalies with large scale datasets enabling the detection of new attack vectors as mentioned above. The uniqueness of this article comes into play because the research group decided to use ChatGPT as a research model and see how well it would perform against another machine learning language. It was a remarkably interesting report, and the reader will have to read the rest of the article if you want to know the results, but the authors will say that some of the promising results that people were looking for are not the results they got. They looked at the already find ML and threw ChatGPT in the same gauntlet and just like most people would expect it was popping up too many false positives. But is this a good thing? I will decide if you'd like to. The authors noticed this article and how recently it was published and decided to see how tools such as ChatGPT can be manipulated and used for things such as this (threat detection).

3.9 Don't Fear the Artificial Intelligence: A Systematic Review of Machine Learning for Prostate Cancer Detection in Pathology

I know what you're thinking this is a crazy one to be putting in a cybersecurity paper. Well, I mean this is crazy interesting! I didn't ever for a moment ever think that our doctors would be getting replaced by a computer I would say that you are utterly crazy and out of your mind. Well, here we are where the world of science and artificial intelligence and machine learning meet. This entire article was to examine if it is possible to detect, if at all, any signs of radiation poisoning as well as certain cancer types. This is especially interesting because using binary to root out if someone has a certain type of cancer is very cool. It was also especially good at spotting differences in size, color, etc. of any cancerous cells. Although this is still in its infancy, we can speculate that there are going to be more advancements with this coming in the future and possibly we could even see ML start helping surgeons make better cuts that they never even saw or considered before. There are endless amounts of opportunity for success with this and little to know down sides. The author believes that with proper funding and a push in the right direction will help to push ML in the medical field in a positive direction.

3.10 Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions

To end for our scholarly annotated bibliographies the authors had to pull out another medical research paper. This medical research paper, however, tries to address certain pitfalls and shortcomings that the authors have been mentioning throughout this entire research paper. This article clearly defines whether there is not a clear data set present, data set characteristics, model design, or clinical implications present in the machine learning and artificial intelligence model. The author of this article aims to provide some background and to finally address all the questions we have been wondering about, what it would look like without a human visit face to face, and we already have those called Teladoc. There are telehealth professionals that do not do anything every day and still get paid to just make sure that the "answering machine" (machine learning model and AI model) is taking all of the patients' needs and meeting them. The authors all agree that this article needs to be given more love and that this needs to be picked up by the medical community.

And this finally concludes our ten scholarly article review!

4. Practitioner Presentations

These practitioner presentations are just as they sound, they are presentations that have been given by industry standard individuals that have practical tools for use in a contained environment. These seminars, presentations, and videos aim to provide a unique perspective from people who are in industry and using these specific technologies and methodologies. These practitioner presentations are intended for academia uses, clients, and other professionals trying to gain insight into ML real-world applications.

4.1 DEF CON 25 - Weaponizing Machine Learning: Humanity Was Overrated Anyway

Has the reader ever thought about what it would be like if we had an AL that can create threats and defensive mechanisms in real-time all by itself? Well look no further because Dan "AltF4" Petro and Ben Morris two of the industry's leading cyber response and machine learning minds come together to bring you Deep Hack. Deep Hack is a conglomeration of hacking tools and generative AI that can not only predict cyber-attacks but can also defend against them if needed. This presentation was given at DEF CON 25 and was short yet very insightful. This presentation was going over real-world application examples with how to not only weaponize AI and machine learning but also how to enact ML proactively defensively to better predict where cyber-attacks are going to be. The uniqueness of this article comes into play where you can take the defensive and try to predict threat vectors as well as vice versa you can take the ML in the defensive position to better adapt to cyber incident response. The authors believe that this forty-five-minute presentation is not only worth the watch, but it provides meaningful takeaways as well. One key takeaway that we are going to leave the reader with it what happens when machine learning has both defensive and offensive techniques? Would these techniques negate each other or would one rule over the other? The author's conclusion to this question is simple, usually attack techniques are more sophisticated than defensive techniques so if there are too many offensive and not enough defensive techniques.

4.2 DEF CON 31 - Growing the Community of AI Hackers w Generative Red Team

Have you ever been interested in just getting into AI and machine learning? Well, this presentation is for you. It culminates aspects from every part of not only cybersecurity, but this brings in great takeaways from AI and ML as well. The main points from this presentation entail the fact that Austin from Seed AI has created more non-for-profit organizations and CTF competitions from different local places in Houston and other areas around the greater TX area that helps youths and those needing the proper helping hand in getting into cyber and AI. They got into dealing with DoD and got an entire group to come together to create more competitions and cash prizes. In the broader sense they get into how anomalies can not only affect the data quality and model accuracy, but it can also impact on the overall health and prediction percentage that the ML can adequately and accurately predict another anomaly or threat. The presenters concluded that we need to keep our thinking hats on and keep an open mind when it comes to different threats and what ML and AI can do to help us facilitate the process of categorizing and effectively managing certain incidents. The authors believe that this is a good step in thinking the right way but in all actuality, you do not know what you don't know, and another zero-day attack could be happening right now, and we wouldn't be the wiser. Hopefully, we can change the course of the academic and professional setting to make sure that people are more open and accepting to changes.

4.3 DEF CON 24 - Machine Duping 101: Pawning Deep Learning Systems

As stated above in the previous papers and presentations, deep learning and machine learning are still developing themselves and still trying to be tapped into. With the presentation has a unique spin on ML and it has a lot to do with automated testing and self-driving cars. ML can start to predict where the road is going to be and make best decisions as to where to move on the road. Take tesla to explain they use extensive amounts of ML and AI to help it along.

4.4 Crowdsource: Applying machine learning to web technical documents to automatically identify malware capabilities

Discussing machine learning -based approach to analyzing and extracting malware is a unique idea. When the authors first saw this topic jumped off the shelf because think of all the cyber incident response algorithms this could be applied to as well how many real-world applications this technology could have. Adding to our existing infrastructure would be a huge advantage that ML and deep learning can take care of in not only real time but also after actions reports or reviews if any shall be conducted. In this article it poses questions and solutions to help aid security professionals as well as quickly pinpointing the area of concern. This is good for not only defensive malware but also offensive. The authors will leave you with that for now.

4.5 From MLOps to MLOops - Exposing the Attack Surface of Machine Learning Platforms

As mentioned in our groups previous research summaries and presentations machine learning is incorporating with everything that it can get its hands on. In this ever-growing landscape as well know there will be dev ops and different sections in a company. When focusing on the operations side you can incorporate these ML languages and models to help further assist and facilitate the building, evaluation, training and sharing of information that is gained by these machine learning models.

5. Conclusion

In conclusion, the fields of machine learning, deep learning, convolutional neural networks, and many more are shaping the way that we view and touch our modern technology. Machine learning is the backbone of the system, deep learning is the enhancement of its capabilities, and CNN's help us to process data like never before. However, as these technologies evolve they too also will challenge us to think in other ways on how to better secure our systems. The future is looking ever brighter with endless possibilities for up-side but also for the potential for the ever more ending dark-side. This is where security professionals get to come in and step up to choose how we use these technologies and what the ultimate capabilities of these technologies and methodologies is/are. It is very crucial the way we interact with these tools and ensure that a sense of responsibility and potential risks/implications of if it broke out into the wrong hands. What do you believe are the stopping capabilities of AI and ML? Are there any, well I'll leave that up for you to decide…

I hope that you have had an amazing semester with the professor and classmates. I really appreciate being able to collaborate with you all and to get to know you better. I could not be more grateful to be in a class and learning environment that allows me to grow and to become a better person and individual.

Thank you once again for everything and if there is anything you will ever need, please do let me know!

6. References

Aaryn Frewing, Alexander B. Gibson, Richard Robertson, Paul M. Urie, Dennis Della Corte; Don't Fear the Artificial Intelligence: A Systematic Review of Machine Learning for Prostate Cancer Detection in Pathology. *Arch Pathol Lab Med* 1 May 2024; 148 (5): 603–612. doi: https://doi.org/10.5858/arpa.2022-0460-RA

Apruzzese, G., Laskov, P., De Oca, E. M., Mallouli, W., Rapa, L. B., Grammatopoulos, A. V., & Di Franco, F. (2022). The role of machine learning in cybersecurity. *Digital Threats Research and Practice*, *4*(1), 1–38. https://doi.org/10.1145/3545574

Black hat. (2024). https://www.blackhat.com/us-24/briefings/schedule/#from-mlops-to-mloops---exposing-the-attack-surface-of-machine-learning-platforms-39309

Cai Y, Gong D, Tang L, Cai Y, Li H, Jing T, Gong M, Hu W, Zhang Z, Zhang X, Zhang G

Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions

J Med Internet Res 2024;26:e47645

URL: https://www.jmir.org/2024/1/e47645

DOI: 10.2196/47645

Detecting insider threats in cybersecurity using machine learning and deep learning techniques. (2023, November 23). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10421133

Harnessing the speed and accuracy of machine learning to advance cybersecurity. (2023, July 24). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10487371/figures#figures

K. Mittal, K. S. Gill, R. Chauhan, K. Joshi and D. Banerjee, "Blockage of Phishing Attacks Through Machine Learning Classification Techniques and Fine Tuning its Accuracy," *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Bangalore, India, 2023, pp. 1-5, doi: 10.1109/SMARTGENCON60755.2023.10442605.

M. Ozkan-Okay *et al.*, "A Comprehensive Survey: Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions," in *IEEE Access*, vol. 12, pp. 12229-12256, 2024, doi: 10.1109/ACCESS.2024.3355547

Nejati, S. (2019). Testing Cyber-Physical Systems via Evolutionary Algorithms and Machine Learning. *2019 IEEE/ACM 12th International Workshop on Search-Based Software Testing (SBST)*, 1. https://doi.org/10.1109/sbst.2019.00008

N. S. Huynh, S. De La Cruz and A. Perez-Pons, "Denial-of Service (DoS) Attack Detection Using Edge Machine Learning," *2023 International Conference on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, 2023, pp. 1741-1745, doi: 10.1109/ICMLA58977.2023.00264.

Rumman Chowdhury, Cattell, S., Chowdhury, R., & Carson, A. (2023). Growing the community of AI hackers with the generative Red Team. *Hack the Future*. https://media.defcon.org/DEF%20CON%2031/DEF%20CON%2031%20presentations/Sven%20Cattell%20Rumman%20Chowdhury%20Austin%20Carson%20-%20Growing%20the%20Community%20of%20AI%20Hackers%20with%20the%20Generative%20Red%20Team.pdf

Sarker, I.H., Kayes, A.S.M., Badsha, S. *et al.* Cybersecurity data science: an overview from machine learning perspective. *J Big Data* **7**, 41 (2020). https://doi.org/10.1186/s40537-020-00318-5

Saxe, J., Turner, R., Blokhin, K., & Nazario, J. (n.d.). CrowdSource: Applying machine learning to web technical documents to automatically identify malware capabilities. *A DARPA Cyber Fast Track Research Effort*. https://media.blackhat.com/us-13/US-13-Saxe-CrowdSource-An-Open-Source-Crowd-Trained-Machine-Learning-Model-Slides.pdf